Concept Decomposition by Fuzzy k-means Algorithm

Jasminka Dobsa
University of Zagreb, Faculty of Organization and Informatics,
42000 Varazdin, Pavlinska 2
jdobsa@foi.hr

Bojana Dalbelo Basic
University of Zagreb, Faculty of Electrical Engineering and Computing,
Unska 3, 10000 Zagreb, Croatia
bojana.dalbelo@fer.hr

Abstract

The method of Latent Semantic Indexing (LSI) is an information retrieval technique using a low-rank singular value decomposition (SVD) of the term-document matrix. Although the LSI method has empirical success, it suffers from the lack of interpretation for the low-rank approximation and, consequently, the lack of controls for accomplishing specific tasks in information retrieval. A method introduced by Dhillon and Modha is an improvement in that direction. It uses centroids of clusters or so called concept decomposition for lowering the rank of the term-document matrix. Our work is focused on improvements of that method using fuzzy k-means algorithm. Also, we compare the precision of information retrieval for the two above methods.

1. Introduction

The task of information retrieval is to extract relevant documents for a certain query from the collection of documents. As large sets of documents are now increasingly common, there is a growing need for fast and efficient for information retrieval algorithms. The algorithms we are dealing with are embedded in the vector space model.

The vector space model is implemented by creating the term-document matrix and a vector of query. Let the list of relevant terms be numerated from 1 to m and documents be numerated from 1 to n. The term-document matrix is an \( m \times n \) matrix \( A = [a_{ij}] \) where \( a_{ij} \) represents the weight of term i in document j. On the other side, we have query or customer’s request. In the vector space model, queries are presented as \( m \)-dimensional vectors. The classic vector space model is based on literal matching of terms in the documents and the queries. But we certainly know that literal matching of terms does not necessarily retrieve all relevant documents. The problem of a synonymy (more words with the same meaning) and a polysemy (a word with multiple meaning) are two major obstacles in the information retrieval. The method of LSI was introduced in 1990 [4] and improved in 1995 [3]. It represents documents as approximations and tends to cluster documents on similar topics even if their term profiles are somewhat different. This approximate representation is accomplished using a low-rank singular value decomposition (SVD) approximation of the term-document matrix. Kolda and O’Leary [8] proposed replacing SVD in LSI by the semi-discrete decomposition that saves memory space. Although the LSI method has empirical success, it suffers from the lack of interpretation for the low-rank approximation and, consequently, the lack of controls for accomplishing specific tasks in information retrieval. Papadimitriou et al. [9] made an effort in the direction of rigorous explanation of empirical success of LSI under certain conditions. Also, they proposed the technique of random projections as a way to speed up LSI. A method by Dhillon and Modha [5] uses centroids of clusters achieved by the spherical k-means algorithm or so-called concept decomposition for lowering the rank of the term-document matrix. Applying this method, the space on which we are projecting the term-document matrix is more interpretable. Namely, it is a space of linear combinations of centroids of clusters. Further, the concept decomposition method is computationally more efficient and requires less memory then LSI. Our work was focused on improvements of this method using the fuzzy k-means algorithm. We have experimentally showed that the projection of the term-document matrix on centroids achieved by the fuzzy k-means algorithm gives better approximation of the term-document matrix in the sense of the Frobenious norm. Also, we investigate how this improvement in approximation reflects on information retrieval. In [5], it is shown experimentally that centroids achieved by the spherical k-means algorithm tend to
orthonormality as \( k \) raises. We will show here that centroids achieved by fuzzy \( k \)-means tend to orthonormality faster.

2. The vector space model and LSI

Let the \( m \times n \) matrix \( A = [a_{ij}] \) be the term-document matrix. Then \( a_{ij} \) is the weight of the \( i \)-th term in document \( j \). We have used \( \tan \) weighing, which means that after creating the term-document matrix based on the frequency of the \( i \)-th term in document \( j \), we have just normalized columns of the matrix to be of unit norm. The term-document matrix has an important property that it is sparse, i.e. most of its elements are zeros.

A query has the same form as a document, it is a vector, which on the \( i \)-th place has the frequency of the \( i \)-th term in the query. Let the query be \( q = [q_i] \). We never normalize the vector of the query because it has no effect on document ranking.

A common measure of similarity between the query and the document is the cosine of the angle between them.

In order to rank documents according to their relevance to the query, we compute

\[
    s = q^T A
\]

where the \( j \)-th entry in \( s \) represents the score in relevance of the \( j \)-th document.

The LSI method is just a variation of the vector space model. The fundamental mathematical result that supports LSI [7] is that for any \( m \times n \) matrix \( A \), the following singular value decomposition exists

\[
    A = U \Sigma V^T
\]

where \( U \) is the \( m \times m \) orthogonal matrix, \( V \) is the \( n \times n \) orthogonal matrix and \( \Sigma \) is \( m \times n \) diagonal matrix

\[
    \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p)
\]

where \( p = \min(m, n) \) and \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0 \). The \( \sigma_i \) are the singular values and \( u_i \) and \( v_i \) are the \( i \)-th left singular vector and the \( i \)-th right singular vector respectively.

The second fundamental result [2,6] is the theorem by Eckart and Young which states that the distance in the Frobenius norm between \( A \) and its rank-\( k \) approximation is minimized by the approximation \( A_k \). Here

\[
    A_k = U_k \Sigma_k V_k^T
\]

where \( U_k \) is the \( m \times k \) matrix whose columns are the first \( k \) columns of \( U \), \( V_k \) is the \( n \times k \) matrix whose columns are the first \( k \) columns of \( V \), and \( \Sigma_k \) is the \( k \times k \) diagonal matrix whose diagonal elements are the \( k \) largest singular values of \( A \).

The ranking of the documents according to their relevance to the query for the LSI method is performed by formula (1) applying rank-\( k \) approximation \( A_k \), instead of the original term-document matrix \( A \).

3. Fuzzy k-means algorithm

The fuzzy \( k \)-means algorithm (FKM) [10] generalizes the classic or hard \( k \)-means algorithm. The goal of \( k \)-means algorithm is to cluster \( n \) objects (here documents) in \( k \) clusters and find \( k \) mean vectors for clusters (here centroids). In the context of the vector space model for information retrieval we call these mean vectors concepts. The spherical \( k \)-means algorithm used in [5] is just a variation of the hard \( k \)-means algorithm which uses the fact that document vectors (and concept vectors) are of the unit norm.

As opposed to the hard \( k \)-means algorithm which allows a document to belong to only one cluster, FKM allows a document to partially belong to multiple clusters. FKM seeks a minimum of a heuristic global cost function

\[
    J_{fuzz} = \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}^b \|x_j - c_i\|^2,
\]

where \( x_j, j = 1, \ldots, n \) are vectors of documents, \( c_i, i = 1, \ldots, k \) are concept vectors, \( \mu_{ij} \) is the fuzzy membership degree of document \( x_j \) in the cluster whose concept is \( c_i \) and \( b \) is a weight exponent in the fuzzy membership. In general, the \( J_{fuzz} \) criterion is minimized when concept \( c_i \) is close to those documents that have high a fuzzy membership degree for cluster \( i, i = 1, \ldots, k \). By solving a system of equations \( \frac{\partial J_{fuzz}}{\partial c_i} = 0 \) and \( \frac{\partial J_{fuzz}}{\partial \mu_{ij}} = 0 \), we achieve stationary point

\[
    \mu_{ij} = \frac{1}{\sum_{i=1}^{k} \left( \frac{\|x_j - c_i\|^2}{\|x_j - c_f\|^2} \right)^{b-1}}, \quad i = 1, \ldots, k; j = 1, \ldots, n.
\]
\[ C_i = \frac{\sum_{j=1}^{n} \mu_y^i x_j}{\sum_{j=1}^{n} \mu_y^i}, \quad i = 1, \ldots, k \]  

(7)

for which the cost function reaches a local minimum.

We will obtain concept vectors starting with arbitrary concept vectors \( C_i^{(0)}, i = 1, \ldots, k \) and computing fuzzy membership degrees \( \mu_y^{(0)} \), cost function \( J_{\text{fuzz}}^{(0)} \) and new concept vectors \( C_i^{(1)} \) iterative, where \( t \) is the index of iteration, until \( |J_{\text{fuzz}}^{(t)} - J_{\text{fuzz}}^{(0)}| < \varepsilon \) for some threshold \( \varepsilon \).

4. Concept decomposition

Our target is to approximate each document vector by a linear combination of concept vectors. The concept matrix as an \( m \times k \) matrix which \( j \)-th column is the concept vector \( c_j \), that is

\[ C_k = [c_1, c_2, \ldots, c_k]. \]  

(8)

If we assume linear independence of the concept vectors, then it follows that the concept matrix has rank \( k \). Now we define concept decomposition \( D_k \) of the term-document matrix \( A \) as the least-squares approximation of \( A \) on the column space of the concept matrix \( C_k \). Concept decomposition is an \( m \times n \) matrix

\[ D_k = C_k Z^* \]  

(9)

where \( Z^* \) is the solution of the least-squares problem, that is

\[ Z^* = (C_k^T C_k)^{-1} C_k^T A. \]  

(10)

It can be shown that for the term-document matrix rank \( k \) approximation achieved by SVD satisfies

\[ A_k = U_k \Sigma_k V_k^T = U_k (U_k^T U_k)^{-1} U_k^T A = U_k U_k^T X. \]  

(11)

So, this approximation is, in fact, the least-squares approximation of matrix \( A \) onto the column space of matrix \( U_k \).

4. Experimental results

Experiments are carried out on standard MEDLINE and CRANFIELD test data sets. Each test data set comes with a collection of documents, a collection of queries and relevance judgments for each query. The relevance judgments are lists of the documents relevant to the specific document. While MEDLINE test data set consists of 1033 documents and 30 queries, CRANFIELD test data set consists of 1400 documents and 225 queries. The list of terms is formed by extracting all terms from the documents and then ejecting terms that occur in only one document and terms on the stop list of 493 common words (SMART list of the stop words). Terms were not stemmed or variations of words were not mapped to the same root form. After this procedure we obtained a list of 5940 terms for MEDLINE test data set and 4758 terms for CRANFIELD test data set.

TEST A. First, we measure the precision of \( k \)-rank approximation \( P_k \) achieved by SVD, concept decomposition by the spherical \( k \)-means algorithm (CDSKM) and concept decomposition by the fuzzy \( k \)-means algorithm (CDFKM) for different ranks of approximation \( k \). A common measure is the fuzzy \( k \)-medoid. A common measure is the Frobenius norm of the difference of the term-document matrix and its approximation \( \| A - P_k \|_F \). From the theorem of Eckard and Young, we know that the best approximation is achieved by SVD. Here, the emphasis is on the comparison of approximations achieved by CDSKM and CDFKM. From Figures 1 and 2, it is clear that the precision of the \( k \)-rank approximation of the term-document matrix is improved by using CDFKM in comparison to that using CDSKM.

TEST B. Second, we investigate how the precision of approximation is reflected on the precision of information retrieval. For this comparison we use the standard measure of mean average precision that measures the average precision on standard recall levels [1]. On Figures 3 and 4, a comparison in performance of the LSI method, CDSKM method and CDFKM method is shown.
Figure 1. Comparison of approx. errors $\|A - P_k\|_F$ for CRANFIELD test data set

Figure 2. Comparison of approx. errors $\|A - P_k\|_F$ for MEDLINE test data set

Figure 3. Mean average precision for CRANFIELD test data set

Figure 4. Mean average precision for MEDLINE test data set

Figure 5. Average scalar product for CRANFIELD test data set

Figure 6. Average scalar product for MEDLINE test data set
Finally, we measure the average inner product between concept vectors $c_j$, $j = 1, \ldots, k$ as

$$\frac{2}{k(k-1)} \sum_{j=1}^{k} \sum_{i \neq j} c_j^T c_i.$$  \hfill (12)

The average inner product takes values in interval $[0,1]$, where smaller values correspond to concept vectors whose average angle between them is closer to $\pi/2$. From Figures 5 and 6, we see that concept vectors achieved by fuzzy $k$-means algorithm tend to orthonormality faster then those achieved by the spherical $k$-means algorithm, especially for the MEDLINE test data set.

5. Conclusion and discussion

Concept decomposition methods are computationally more efficient and require less memory then SVD. Further, they can exploit the sparsity of the term-document matrix. The computational complexity of the spherical $k$-means and of fuzzy $k$-means is $O(nmkT)$, where $n$ is the number of documents, $m$ is the number of terms, $k$ is the number of clusters and $T$ is the number of iterations. Although complexity for these two algorithms is the same, fuzzy $k$-means is much more time consuming, partly due to more computational operations, partly due to slower convergence. We suggest here a modification of the fuzzy $k$-means algorithm in such a way that the fuzzy membership degree of document $x_i$ is calculated only for those clusters whose concept vectors are closest to the document vector $x_i$ (fuzzy membership degrees for other clusters should be 0).

The fact that matrix approximations achieved by CDSKM and CDFKM are worse than by SVD does not reflect on the precision of information retrieval for the two standard test data sets we have used. We see that the mean average precision is comparable for CDSKM and CDFKM with LSI method for the CRANFIELD test data set and that CDFKM even outperforms LSI in the case of MEDLINE test data set. Also, we notice that, for low ranks of approximation, mean average precision grows fastest for CDFKM.

The reason LSI method performs better compared to the term-matching vector space method is the fact that LSI clusters documents on similar topics, projecting them on the first $k$ left singular vectors, thus enabling better treatment of the problem of synonyms. The reason CDSKM and CDFKM methods perform well is the same.

Generally, mean average precision grows with rank of approximation, but we can see that for the LSI method growth is more stable then for CDSKM and CDFKM. For CDSKM and CDFKM we can notice drops of mean average precision for some ranks of approximation. In future work, the dependence of mean average precision and cluster validity could be investigated in order to determine the best number of clusters for a given data set and connect it with the choice of the rank of approximation in concrete application.

6. References


